

Análise de Relações de Coautoria nos Programas de Pós-Graduação em Ciência da Computação no Brasil

Jeojildo Pereira¹, Helder Matos¹, Jean Carlos Costa¹, Reginaldo Santos¹

¹ Laboratório de Inteligência de Dados (LID)
Programa de Pós-Graduação em Computação (PPGCOMP)
Universidade Federal do Pará (UFPA)

{jeojildo.pereira, helder.matos, jean.carvalho.costa}@icen.ufpa.br
regicsf@ufpa.br

Abstract. *This study investigates the structure of scientific collaboration in Brazilian graduate programs in Computer Science from 2014 to 2023. Co-authorship relationships are modeled as a network to identify collaboration patterns. Using the KDD process, the analysis uses data extracted from CAPES Lattes platform, covering 1,757 permanent faculty members across 67 programs and 63 universities. Community detection is conducted through modularity algorithm, while weighted degree and betweenness centrality measure structural relevance. The results show cohesive communities, key co-authors as hubs and bridges, and influential programs, providing a quantitative view of collaboration, co-authorship and network structure.*

Resumo. *Este estudo investiga a estrutura de colaboração científica nos programas de pós-graduação em Ciência da Computação no Brasil de 2014 a 2023. As relações de coautoria são modeladas como uma rede para identificar padrões de colaboração. Com o processo KDD, a análise emprega dados extraídos da plataforma Lattes da CAPES, abrangendo 1.757 docentes permanentes de 67 programas em 63 universidades. A detecção de comunidades usa algoritmo de modularidade, enquanto centralidades de grau e de intermediação medem relevância estrutural. Os resultados mostram comunidades coesas, coautores-chave atuando como hubs e bridges, e programas influentes, fornecendo visão quantitativa de colaboração e da rede.*

1. Introdução

A colaboração científica constitui elemento central na consolidação e expansão do conhecimento, especialmente na área de Ciência da Computação, cuja produção é fortemente estruturada em redes de coautoria entre pesquisadores e instituições. As relações de coautoria permitem modelar interações acadêmicas como redes complexas, possibilitando a análise estrutural de fluxos de conhecimento, padrões colaborativos e articulações interinstitucionais [Newman 2001]. No Brasil, os programas de pós-graduação em Ciência da Computação (PPGCC) exercem função estratégica na formação de pesquisadores e na produção científica nacional. Nesse contexto, compreender a estrutura das redes de colaboração torna-se relevante para análises quantitativas sobre conectividade científica e cooperação interinstitucional.

Estudos anteriores investigaram métricas de colaboração e ranqueamento de autores em redes acadêmicas [Digiampietri et al. 2012, Lima et al. 2024, Costa et al. 2022].

Entretanto, ainda são limitados os estudos que analisam, em escala nacional, a organização comunitária das redes de coautoria dos PPGCCs brasileiros por meio de algoritmos formais de detecção de comunidades e métricas estruturais de redes complexas. Este trabalho propõe uma análise estrutural envolvendo 1.757 docentes permanentes distribuídos em 67 PPGCCs no Brasil. A metodologia fundamenta-se no processo de *Knowledge Discovery in Databases* (KDD), incorporando modelagem em grafos ponderados e aplicação dos algoritmos *Louvain*, *Leiden* e *Greedy Modularity* para identificação de comunidades estruturais, e são calculadas métricas de centralidade de grau ponderada e centralidade de intermediação, permitindo identificar atores estratégicos e padrões de intermediação na rede.

2. Trabalhos Relacionados

Diversos estudos investigaram relações de coautoria utilizando métricas estruturais e técnicas de mineração em redes científicas. [Digiampietri et al. 2012] analisaram redes de coautoria utilizando *Author Rank* e distância de Levenshtein com heurísticas de resolução de entidades. [da Silva et al. 2020] combinaram métricas estruturais com técnicas de aprendizado de máquina baseadas em *Random Forest*. Em âmbito internacional, [Peng et al. 2013] exploraram métricas de centralidade e algoritmos de mineração de comunidades para identificação de atores estratégicos em redes científicas.

Outros trabalhos aplicaram algoritmos de detecção de comunidades em redes de coautoria. [Lima et al. 2024] utilizaram o coeficiente de similaridade *Szymkiewicz-Simpson* em conjunto com o algoritmo de *Louvain*, enquanto [Tsolakidis et al. 2012] investigaram características estruturais de redes de coautoria combinadas com detecção de comunidades baseada em *Louvain*. Apesar desses avanços, ainda permanece pouco explorada a estrutura comunitária das redes de coautoria dos PPGCC brasileiros. Este trabalho utiliza o processo KDD com dados da plataforma Lattes, integrando grau ponderado e centralidade de intermediação e o algoritmo de *Leiden* para analisar as relações de coautoria dos PPGCC no Brasil do período de 2014 a 2023.

3. Metodologia

A metodologia adotada fundamenta-se no processo KDD. Nesta pesquisa, as etapas do KDD foram implementadas para estruturar a preparação dos dados, a construção da rede de coautoria e a identificação de padrões estruturais de colaboração científica. A Figura 1 ilustra o processo KDD adotado, cujo código está disponível no GitHub¹. O método compreende quatro etapas: seleção de dados (PPGCC e docentes), pré-processamento

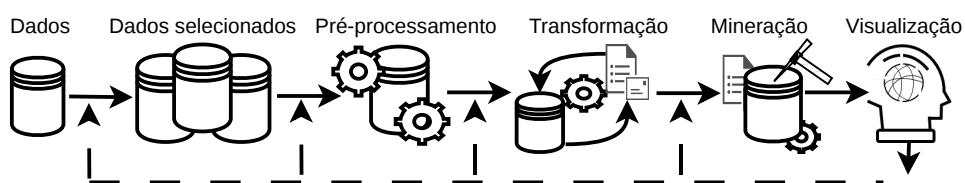


Figura 1. Implementação do processo de KDD nos relações de coautoria.

¹Link para o repositório: <https://github.com/jeojildo/ppgcc-coautorias>

dos currículos Lattes, construção da rede de coautoria ponderada e análise estrutural para identificar padrões de colaboração. Nesse contexto, a análise considera pesquisadores dos PPGs em Ciência da Computação e suas coautorias. Os pesquisadores são vértices, e as coautorias formam as arestas da rede.

3.1. Seleção dos dados

O processo de seleção dos dados iniciou em maio de 2024, com foco no cenário recente da pós-graduação *stricto sensu* em Ciência da Computação no Brasil. Os dados foram obtidos das plataformas CAPES e Sucupira (período 2021–2024), incluindo informações sobre programas, docentes, localização e áreas. Foram selecionados 67 PPGs, distribuídos em 63 instituições, totalizando 1.757 docentes. Os currículos Lattes foram extraídos em XML via ferramenta automatizada, a partir do ID CNPq. Os dados contemplam informações dos pesquisadores e de suas publicações, permitindo mapear relações de coautoria com base nas listas de autores.

3.2. Pré-processamento e Transformação dos Dados

Os dados foram restritos ao período de 2014 a 2023, considerando apenas artigos completos publicados em conferências e periódicos. Após a remoção de duplicidades, inconsistências e registros incompletos, realizou-se a padronização das informações institucionais e a desambiguação de autores com base no ID CNPq, assegurando a unicidade dos nós. O conjunto final compreende 230.917 publicações, nas quais cada par de autores define uma relação de coautoria. Em seguida, os dados foram convertidos em uma rede de coautoria ponderada, representada por uma matriz de adjacência, na qual autores correspondem a nós, colaborações a arestas, e os pesos indicam a frequência de coautorias, capturando a intensidade das relações.

3.3. Mineração de dados

Esta etapa consiste na identificação de padrões na rede de coautoria por meio de análise de centralidade de intermediação, a fim de identificar autores que atuam como pontes entre diferentes regiões da rede. Para a detecção de comunidades, foram avaliados os algoritmos *Louvain* [Blondel et al. 2008], *Leiden* [Traag et al. 2019] e *Greedy Modularity* [Clauset et al. 2004], sendo o *Leiden* adotado por apresentar melhor equilíbrio entre modularidade e coesão das comunidades. A qualidade das comunidades é avaliada pela modularidade, permitindo caracterizar a estrutura comunitária da rede de coautoria.

4. Resultados

Esta seção apresenta os resultados das relações de coautoria e dos padrões de colaboração entre pesquisadores dos PPGs no Brasil. A análise utiliza uma matriz de coautorias construída na etapa de transformação dos dados². Dos 1.757 pesquisadores iniciais, 32 foram excluídos por não possuírem publicações em coautoria, resultando em 1.725 pesquisadores na matriz final.

²Link para o arquivo CSV da matriz de coautorias: <https://bit.ly/4bokIDq>

4.1. Centralidade de intermediação

A Figura 2 mostra uma correlação positiva e moderada entre a centralidade de intermediação e o grau ponderado. Esse padrão indica que autores com valores elevados em ambas as métricas atuam como pontes entre comunidades e sustentam a conectividade global da rede. Por outro lado, autores com alto grau ponderado e intermediação moderada formam concentrações locais, com menor influência no fluxo entre comunidades. Essa distinção orienta a análise da seção seguinte, na qual esses perfis são examinados em conjunto com a estrutura de comunidades.

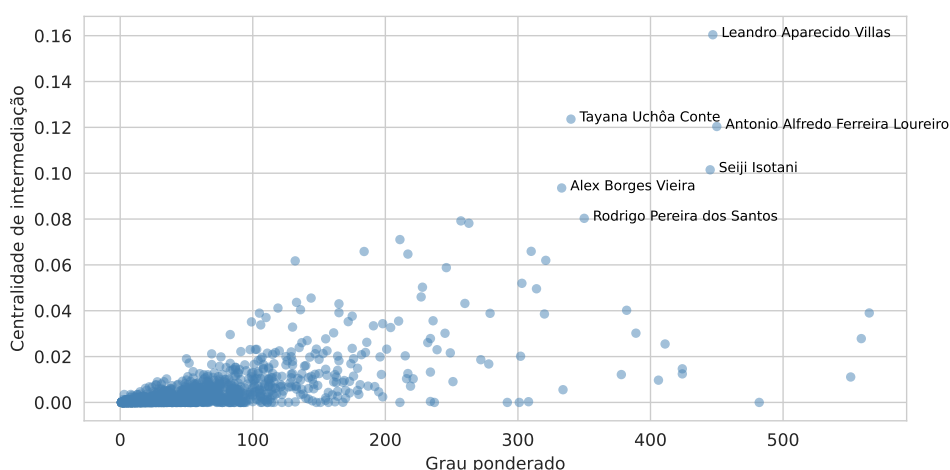


Figura 2. Relações entre centralidade intermediação e graus ponderados (Pearson $r = 0,666$, $p = 1,02 \times 10^{-159}$; Spearman $\rho = 0,684$, $p = 2,66 \times 10^{-171}$).

4.2. Detecção de comunidades

O algoritmo *Leiden* foi aplicado com resolução de 0,6, identificando 24 comunidades em uma rede de coautoria com 1.725 autores e 9.415 arestas. A baixa densidade (0,0063) indica uma estrutura esparsa, enquanto a alta modularidade (0,773) revela forte organização comunitária, com predominância de colaborações internas. A maior comunidade possui 239 autores; a média é de 69 e a mediana de 40. Além disso, 85,8% do peso das arestas corresponde a relações intracomunitárias, frente a 14,2% intercomunitárias, o que sugere colaboração concentrada em grupos coesos, possivelmente associados a proximidade institucional ou grupos de pesquisa consolidados.

O grau de integração entre comunidades pode ser analisado pela participação intercomunitária (Tabela 1). Há variação entre comunidades mais abertas, como a 7 (participação externa = 0,322), e mais fechadas, como a 1 (0,217). As principais comunidades se conectam a um número de grupos parceiros, entre 15 e 18. Esse padrão indica que, apesar da forte coesão interna, a rede não forma blocos isolados. A diversidade das conexões externas, medida pela entropia (entre 5,349 e 6,726), sugere distribuição das colaborações entre múltiplas comunidades, sem concentração em um único parceiro.

O perfil institucional das comunidades reforça a interpretação de grupos multi-institucionais. As comunidades com maior diversidade institucional incluem a comunidade 0, com 48 instituições (entropia institucional de 3,557), e a comunidade 2, com 34 instituições (entropia de 2,801). Também se destacam a comunidade 7 e 1, com 30

Tabela 1. Perfil estrutural por comunidade.

Comunidade	Nº de Autores	Peso das coautorias internas	Participação externa	Entropia	Nº de Instituições	Instituição Principal	Dominância (%)	Entropia Institucional
0	217	5167	0,298	6,726	48	UFAM	9,22	3,557
1	185	3828	0,217	6,840	30	UFF	23,24	2,792
2	151	4900	0,252	6,419	34	UFMG	31,13	2,801
3	145	2603	0,287	5,977	21	UFU	18,62	2,607
4	145	2197	0,310	5,987	25	UNICAMP	22,76	2,840
5	131	1985	0,285	5,877	26	USP-SC	33,59	2,421
6	131	3006	0,290	6,330	21	UFPE	25,19	2,526
7	116	2362	0,322	5,980	30	UFRGS	20,69	2,862
8	99	1492	0,268	5,349	14	UFPE	23,23	2,237

instituições. Em geral, a instituição mais frequente em cada comunidade concentra apenas uma fração limitada dos autores: nas maiores comunidades, a dominância varia de 9,22% (comunidade 0, UFAM) a 31,13% (comunidade 2, UFMG). Esses resultados indicam composição institucional heterogênea, sem predominância absoluta de uma única instituição na maior parte das comunidades analisadas. Em conjunto com os indicadores estruturais, esse padrão sugere comunidades coesas internamente e, ao mesmo tempo, abertas à interação entre grupos.

A Tabela 2 apresenta as 20 instituições com maior volume de colaboração na rede. A UFMG lidera em força total com 5.582, seguida por UFPEL com 4571 e USP-SC com 4528. Em termos de intermediação, destacam-se UFJF (0,0310), UNESP-SJRP (0,0291) e UFPE (0,028), indicando papel de ponte na conectividade global da rede. Quanto à abertura externa, UFRJ-COMP (72,25%), PUC-RIO (64,29%) e UFAL (57,93%) apresentam a maior proporção de colaborações fora de sua própria instituição, sugerindo perfil de maior diversificação colaborativa.

Tabela 2. Instituições com maior volume de colaboração na rede de coautoria.

Instituição	Força Total	Intermediação	Abertura Externa	Instituição	Força Total	Intermediação	Abertura Externa
UFMG	5582	0,0013	42,17%	UNISINOS	2082	0,0074	25,84%
UFPEL	4571	0,0007	15,55%	UFJF	2073	0,0310	28,32%
USP-SC	4528	0,0139	53,09%	UFPA	2029	0,0188	43,72%
UFPE	3964	0,0282	52,02%	UFG	2002	0,0153	34,57%
UFF	3866	0,0243	36,83%	FURG	1898	0,0122	24,24%
UFRGS	3775	0,0237	40,34%	UFPR	1891	0,0182	50,61%
UFC	3258	0,0130	22,59%	UFRJ-COMP	1881	0,0164	72,25%
UNICAMP	2727	0,0001	52,33%	UNESP-SJRP	1810	0,0291	32,93%
UFMA	2663	0,0150	24,37%	UFCG	1748	0,0072	29,52%
UFAL	2092	0,0256	57,93%	PUC-RIO	1630	0,0140	64,29%

Os resultados indicam que a rede de coautoria dos PPGs em Ciência da Computação no Brasil apresenta estrutura esparsa, porém conectada, com evidências do fenômeno de *small world*. A colaboração concentra-se principalmente em grupos recorrentes, enquanto um número reduzido de autores e instituições atua como intermediador entre comunidades. A rede organiza-se em comunidades com alta coesão interna e níveis moderados de colaboração interinstitucional, evidenciando um padrão de produção científica estruturado em grupos colaborativos interligados.

5. Conclusões

Os resultados indicam que as redes de coautoria dos PPGCC no Brasil apresentam estrutura esparsa, porém altamente conectada, com forte organização comunitária e

colaboração concentrada em grupos de maior produtividade. Autores com maior centralidade de intermediação desempenham papel fundamental na conectividade entre comunidades, evidenciando que o maior volume de colaboração está associado a posições estruturais estratégicas e reforçando a importância de *hubs* e atores intermediadores para a manutenção da conectividade global da rede. A detecção de comunidades por meio do algoritmo de *Leiden* revelou grupos coesos, influenciados por fatores institucionais, permitindo identificar estruturas colaborativas bem definidas. Além disso, a abordagem baseada no processo de KDD contribuiu para a organização sistemática da análise e para a robustez dos resultados. Como trabalhos futuros, recomenda-se integrar outras fontes de dados, como ORCID, Google Scholar e OpenAlex, realizar análises temporais para aprofundar a compreensão da evolução das redes e incluir a classificação Qualis CAPES, a fim de correlacionar a estrutura das redes à qualidade da produção científica dos programas.

Referências

- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6).
- Costa, A. R., Nunes, V. T., and Ralha, C. G. (2022). Scientific Collaboration Network Views: A Brazilian Computer Science Graduate Programs Case. *Journal of Information and Data Management*, 13(4).
- da Silva, A. J. N., Breve, M. M., Mena-Chalco, J. P., and Lopes, F. M. (2020). Analysis of co-authorship networks among Brazilian graduate programs in computer science. *CoRR*, abs/2012.12439.
- Digiampietri, L. A., Mena-Chalco, J. P., Silva, G. S., Oliveira, L. B., Malheiros, A. P., and Meira, D. (2012). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil.
- Lima, F., Miranda, L., Vasiljevic, G., and Baranauskas, M. C. (2024). An Analysis of the Authorship and Co-authorship Networks of the Brazilian Human-Computer Interaction Conference. *Journal on Interactive Systems*, 15:265–293.
- Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2).
- Peng, T., Zhang, D., Liu, X., Wang, S., and Zuo, W. (2013). Central Author Mining from Co-authorship Network. In *Proceedings of the 2013 Sixth International Symposium on Computational Intelligence and Design - Volume 01*, ISCID '13, page 228–232, USA. IEEE Computer Society.
- Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).
- Tsolakidis, A., Sgouroplou, C., Papageorgiou, E., Terraz, O., and Miaoulis, G. (2012). Co-authorship Networks in Academic Research Communities: The Role of Network Strength. In *2012 16th Panhellenic Conference on Informatics*, pages 150–155.